AD-A104 041 TEXAS CHRISTIAN UNIV FORT WORTH INST OF BEHAVIORAL R--ETC F/G 5/11
ESTIMATING INTERRATER RELIABILITY IN INCOMPLETE DESIGNS.(U)
AUG 81 L R JAMES, G WOLF, R G DEMAREE N00014-80-C-0315
UNCLASSIFIED IBR-81-14 NL

1 OF 1
AD
A104041

END
DATE
FILMED
10-81
DTIC

12

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
IBR 81-14

2. GOVT ACCESSION NO.
AD-A104 041

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Estimating Interrater Reliability in Incomplete Designs

5. TYPE OF REPORT & PERIOD COVERED
Technical Report,

6. PERFORMING ORG. REPORT NUMBER
IBR-81-14

7. AUTHOR(s)
Lawrence R. James
Gerrit Wolf
Robert G. Demaree

8. CONTRACT OR GRANT NUMBER(s)
N00014-80-C-0315

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Institute of Behavioral Research
Texas Christian University
Fort Worth, Texas 76129

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
NR170-904

11. CONTROLLING OFFICE NAME AND ADDRESS
Organizational Effectiveness Research Programs
Office of Naval Research (Code 452)
Arlington, Virginia 22217

12. REPORT DATE
Aug[redacted]81

13. NUMBER OF PAGES
32

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Agreement
Incomplete Design
Interrater Reliability
Intraclass Correlation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Estimates of interrater reliability are often needed for incomplete designs in which raters (e.g., employees) are nested within targets (e.g., organizations). It is shown that the popular use of estimates based on between-group ANOVAs accompanied by intraclass correlations can be seriously misleading if low variation exists among target means. An alternative based on a within-group procedure is proposed and

9. School of Psychology
Georgia Institute of Technology
Atlanta, Georgia 30332

20. shown to be superior to the intraclass correlation in the condition of low variation among group means, accompanied by low within-group variation.

Accession For	
NTIS GRA&I	☒
DTIC TAB	□
Unannounced	□
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A	

Estimating Interrater Reliability in Incomplete Designs

Lawrence R. James

Georgia Institute of Technology

Gerrit Wolf

University of Arizona

and

Robert G. Demaree

Institute of Behavioral Research

Texas Christian University

Estimating Interrater Reliability in Incomplete Designs

The use of various forms of the intraclass correlation coefficient to estimate interrater reliability has been addressed in a number of recent articles (Bartko, 1976; Bintig, 1980; Fleiss & Shrout, 1978; Kraemer & Korner, 1976; Saal, Downey, & Lahey, 1980; Shrout & Fleiss, 1979). While these articles have focused on "complete designs", where each target is rated by each judge on one or more dimensions (variables), several have touched on "incomplete designs" in which each of $\underline{K}$ targets is rated by a different set of judges (i.e., judges are nested within targets) using the same rating variable(s). A form of intraclass correlation may be used to provide consistent estimates of interrater reliability for incomplete designs, the typical question being whether the judges within each of the $\underline{K}$ targets agreed with respect to their ratings (cf. Ebel, 1951; Guilford, 1954; Shrout & Fleiss, 1979; Winer, 1971).

The incomplete design is employed frequently in areas such as climate research, where $\underline{n_k}$ employees nested in each of $\underline{K}$ ($\underline{k}$=1,...,$\underline{K}$) organizations report perceptions on a climate variable such as "managerial support" (cf. Insel & Moos, 1974; James & Jones, 1974). An interrater reliability is computed from information furnished by a random effects, one-way ANOVA. That is, each of the $\underline{K}$ targets (organizations) assumes the role of a treatment, and ratings (perceptions) provided by the $\underline{n_k}$ judges (employees) furnish values on the dependent variable "X". The n_k need not be equal. A one-way ANOVA is conducted, where a significant F suggests that variation among the scores on X was associated more with differences among targets than with pooled differences among judges nested within targets. An estimate of interrater reliability is obtained by the following equation (cf. Ebel, 1951).

$$\mathit{ICC} = \frac{\mathit{MSB} - \mathit{MSW}}{\mathit{MSB} + (\bar{n}_k - 1)\ \mathit{MSW}} \tag{1}$$

where ICC is an intraclass correlation, MSB is the mean square for between-groups (targets), MSW is the within-group mean square, and $\bar{n}_k$ is a harmonic mean based on the number of judges per group k. The more convential term "group" refers to all judges, or raters, who rated the same target.

Interrater reliability is viewed here as a function of the degree to which raters who rated the same target agreed with respect to their ratings; that is, high interrater reliability is indicated by high within-group agreement, and low interrater reliability is indicated by lack of within-group agreement. The terms interrater reliability and agreement are used interchangeably.

The initial objective of this report is to demonstrate that the ICC above may provide a seriously misleading indicator of interrater reliability. First, inspection of Eq. 1 demonstrates that as MSW decreases, the ICC increases. Thus, the

ICC is a function, in part, of the extent to which within-group agreement is present, as shown by the degree to which raters within each group give the same ratings (Bartko, 1976). Note, however, that the ICC and MSW are based on pooled data, and the ICC estimate applies to all groups. If the separate within-group variances are not homogeneous, then the ICC may overestimate agreement for some groups, estimate it accurately for others, and underestimate it for still others. This potential problem may be checked empirically by a homogeneity of variance test. Suppose, however, that the null hypothesis of equal variances is rejected. Does this suggest that interrater reliability cannot be estimated? Certainly not; it suggests only that a between-group design should not be used and that a separate estimate of agreement should be obtained for each group.

A second and more important point is that even with high agreement among raters in each group, the ICC may be very low. Consider, for example, a scenario in which (a) the raters in each one of K groups responded almost exactly the same, which denotes close to perfect agreement among the ratings for each target and a low MSW; and (b) the mean scores for all K groups were essentially identical, in which case MSB is zero, or approximately so. Inspection of Eq. 1 demonstrates that, given these conditions, the ICC would be equal to zero, or even negative in value.

To illustrate, consider the data presented in Table 1. The data consist of scores on a random variable X, which has five discrete, equally spaced alternatives, for 20 individuals in each of two groups. In accordance with assumptions underlying the use of ANOVA and the _ICC_ (cf. Shrout & Fleiss, 1979; Winer, 1971), it is presumed that (a) groups (e.g., organizations) and raters (e.g., employees) were randomly sampled from populations to which inferences regarding groups and raters are to be made, (b) raters rated independently, (c) the within-group residual components are independently and normally distributed in the population, and (d) the within-group variances are homogeneous. The response frequencies in Table 1 indicate that individuals in each group tended to agree. Agreement is also reflected by the small within-group variances (.211 and .261). However, not only is the _F_-test nonsignificant ($p > .05$), but the _ICC_ is -.047, which is regarded as .00 (Bartko, 1976). This low and obviously misleading _ICC_ is attributed to the essential absence of variation among the group mean ratings (3.00 and 3.05).

Insert Table 1 about here

We hasten to note that lack of variation among group mean ratings does not automatically indicate a misleading _ICC_. For example, low variation among means accompanied by high variation among ratings within groups provides an accurate _ICC_ of approximately zero. Our concern is limited to conditions of the type displayed in Table 1, where low variation among group mean ratings accompanied by low within-group variance results in inaccurate estimates of agreement. Moreover, we submit that such conditions are neither unrealistic nor trivial. Consider, for example, a study in which a different sample of $\underline{n}_k$ inspectors (raters) rates each of $\underline{K}$ airplanes (groups, targets), selected randomly from those airplanes owned by a particular airlines company. It is not unreasonable to assume that (a) this company has followed rigorous maintenance standards in the interest of satisfying safety criteria, and (b) the $\underline{n}_k$ raters of each airplane rate that airplane highly in regard to safety. Are we to conclude that, based on Eq. 1, the inspectors failed to agree with respect to the safety of the airplanes?

In summary, an ICC based on the one-way, between-group ANOVA design has potentially serious deficiencies as an estimator of interrater reliability/agreement. The approach should not be used if group variances are heterogeneous. Moreover, given homogeneity of variance, the ICC will be misleading if group mean ratings do not vary and low variation exists among ratings within groups. Given either of these situaitons, a different method is needed to estimate interrater reliability/agreement. The second objective of this paper is to propose such a method.

Estimating Interrater Reliability Using a Within-Group Design

The proposed method for estimating interrater reliability is based on a within-group design. A within-group design was selected because a separate estimate of agreement may be obtained for each group in an incomplete design, and, of major importance, this estimate is not affected by either failure to have large between-group differences or lack of homogeneity of within-group variances. Furthermore, the estimates for each group may be averaged to furnish an overall estimate of agreement for all groups if the homogeneity of variance assumption is satisfied. As we shall demonstrate, this average may be substantially higher and an obviously more accurate estimator of interrater reliability than the ICC in the condition of major concern (i.e., high within-group agreement and low between-group variation).

Within-group approaches for estimating interrater reliability are not new (Bintig, 1980; Finn, 1970; Selvage, 1976). However, some, but not all, methods and logical principles presented here differ from those of earlier treatments. In addition, Cooper (1976) and Hsu (1979) presented exact small sample and approximate large sample tests designed to ascertain if raters within a group agreed significantly with respect to their ratings on a single Likert-type item. These articles dealt only with significance tests and did not furnish a basis for estimating an interrater reliability coefficient. The present authors are in agreement with Cohen (1960), who expressed the opinion that when reliability is of concern, significance is a trivial point because "one usually expects much more than this [i.e., significance] in the way of reliability in psychological measurement" (p. 44). Consequently, we have devoted our attention to point estimates of interrater reliability and do not consider significance tests.

The presentation of the estimating procedure begins with the variance on a rating variable "$\underline{X}$" in one group, or $\underline{s}_X^2$. For example, in Table 1, $\underline{s}_X^2 = .211$ for Group 1. An $\underline{s}_X^2 = 0$ indicates perfect agreement. Typically, however, $\underline{s}_X^2 \neq 0$, in which case the question is the degree to which raters in the group agreed with respect to their ratings. To develop a statistic that estimates degree of agreement, it is necessary to have a standard or benchmark to compare to $\underline{s}_X^2$. Inasmuch as $\underline{s}_X^2 > 0$ reflects departure from perfect agreement, we shall adopt a benchmark that reflects the expected value of $\underline{s}_X^2$ in a condition of absence of agreement. This expected variance is referred to as "$\underline{\sigma}_E^2$".

Procedures for determining a value of $\underline{\sigma}_E^2$ for a single Likert-type scale in a single group have been presented by Cooper (1976), Finn (1970), Hsu (1979), and Selvage (1976). Finn and Cooper argued that an interrater reliability of zero occurs if raters responded randomly to an item. Random responding implies that each alternative on the rating scale has an equal likelihood of response. Equal likelihood of response, combined with assumptions that (a) raters responded independently and (b) the item $\underline{X}$ is a discrete random variable with multiple alternatives arranged on an interval scale, suggests that $\underline{\sigma}_E^2$ may be calculated using the equation for the variance of discrete, uniform distribution. Specifically, define the item $\underline{X}$ as a random variable which assumes $\underline{A}$ ($\underline{a}$=1,...,$\underline{A}$) finite, equally spaced alternatives (i.e., $\underline{A}$ corresponds to the number of alternatives on $\underline{X}$). Equal likelihood of response connotes that each value of $\underline{A}$ has the same probability of occurrence, or $\underline{P}(\underline{X} = \underline{a}) = 1/\underline{A}$. In other words, the distribution of $\underline{X}$ is uniform or rectangular. As shown in a number of statistical texts (cf. Mood, Graybill,& Boes, 1974), the expected variance of $\underline{X}$ is then:

$$\begin{aligned} \mathrm{Var}(\underline{X}) &= \underline{E}(\underline{X}^2) - [\underline{E}(\underline{X})]^2 \\ &= (\underline{A}^2 - 1)/12 \end{aligned} \qquad (2)$$

Equation 2 provides the desired value of $\underline{\sigma_E}^2$ and is interpreted as the expected variance of $\underline{X}$ associated with equal likelihood of response and zero interrater reliability. A critical point to be made about $\underline{\sigma_E}^2$ is that it is a benchmark connoting an absence of agreement and is to be viewed as a statistical abstraction. Whether raters would ever respond to an item in a sheerly random fashion has no bearing on the appropriateness of relying upon hypothetical random responding as a statistical referent for assessing the extent to which a set of actual responses resemble a set of random responses.

The benchmark $\underline{\sigma_E}^2$ is now employed to estimate the variance in ratings due to nonerror variance and then interrater reliability. Consider first that an observed score on $\underline{X}$, designated $\underline{X_i}$ ($\underline{i}$=1,...,$\underline{n_k}$ subjects) may be represented as $\underline{X_i} = \underline{\mu} + (\overline{\underline{X}} - \underline{\mu}) + \underline{e_i}$, where $\underline{\mu}$ and $\overline{\underline{X}}$ are the population and sample means on tne item, respectively, and $\underline{e_i}$ is an error of measurement. The variance of the $\underline{X_i}$, or $\underline{s_X}^2$, in a sample arises only from variation in the $\underline{e_i}$, and thus $\underline{s_X}^2$ is referred to as "error variance".

If the $\underline{X_i}$ are reflective solely of $\underline{\mu}$, and thus are entirely devoid of error variance, then $s_X^2 = 0$. On the other hand, if the X_i are a function of error exclusively and conform to equal likelihood, random responses, then $s_X^2 = \sigma_E^2$. This suggests that the extent to which the X_i are actually reflective of μ, and may be said to reveal nonerror or true variance, is indicated by $\sigma_E^2 - s_X^2$. The use of $\sigma_E^2 - s_X^2$ to estimate true variance is a heuristic designed to "breakout" of a closed system in which restrictions in variances preclude the use of traditional statistical procedures. Thus, for example, $s_X^2 = 0$ implies that the X_i are solely a function of μ, which is indicated by setting true variance equal to $\sigma_E^2 - 0 = \sigma_E^2$. There is, of course, no such variance (i.e., μ is a constant), but the heuristic shifts the basis of analysis to a different logical system, based on σ_E^2, in which it is possible to estimate interrater reliability.

An estimate of interrater reliability is obtained by placing the estimates of the variances in the equation: (true variance)/(true variance + error variance), or $(\sigma_E^2 - s_X^2)/[(\sigma_E^2 - s_X^2) + s_X^2] = (\sigma_E^2 - s_X^2)/\sigma_E^2$. This equation reduces to the equation suggested by Finn (1970, p. 72), namely $1 - (s_X^2/\sigma_E^2)$, where (s_X^2/σ_E^2) estimates the "proportion of random or error variance present in the observed ratings", and $1 - (s_X^2/\sigma_E^2)$ "gives the proportion of non-error variance in the ratings, a reliability coefficient."

To summarize, the equation for estimating interrater reliability/agreement is:

$$r_{WG} = (\sigma_E^2 - s_X^2)/\sigma_E^2$$

$$= 1 - (s_X^2/\sigma_E^2) \qquad (3)$$

where:

r_{WG} = within-group interrater reliability for a single group of raters who have rated the same target on one discrete, equal interval variable,

s_X^2 = the observed (errôr) variance on variable X for the n_k raters in group k,

σ_E^2 = the variance on X that would be expected if the raters responded randomly, which is estimated by $(A^2 - 1)/12$ for a discrete, uniform distribution (see Eq. 2).

Note that perfect interrater reliability/agreement is indicated by $s_X^2 = 0$, in which case $r_{WG} = 1.0$. Conversely, equal likelihood of response connotes zero reliability and no agreement, in which case $s_X^2 \cong \sigma_E^2$ and $r_{WG} \cong 0$. Given the usual condition in which $0 < s_X^2 < \sigma_E^2$, as s_X^2 approaches σ_E^2, agreement decreases; or, as s_X^2 becomes progressively smaller than σ_E^2, agreement increases.

The use of Eq. 3 is illustrated by application to the data in Table 1. With $A = 5$ (i.e., the item has five alternatives), $\sigma_E^2 = 2.0$ in each of the two groups $[(5^2 - 1)/12]$. Inserting the values of σ_E^2 and the observed variances (s_X^2) into Eq. 3 supplies the desired estimates of r_{WG}; r_{WG} for Group 1 is .89 [i.e., 1 - (.211/2)], and r_{WG} for Group 2 is .87 [i.e., 1 -

(.261/2)]. Clearly, values of .89 and .87 are different than the intraclass correlation of .00, and it is just as clear that the former values are more consistent with the data than the latter value. Furthermore, given the similarity of the two values of $\underline{r_{WG}}$, it is possible to average the values and obtain an overall estimate of interrater reliability for both groups. Averaging is not recommended if the values of $\underline{r_{WG}}$ are dissimilar for the obvious reason that the average would be misleading for at least some groups. A homogeneity of variance test on observed variances might be used to decide whether to average the coefficients across all groups, or perhaps subsets of groups. [Given homogeneity of variance, the average $\underline{r_{WG}}$ may be estimated by 1 - ($\underline{MSW}/\sigma_{\underline{E}}^2$)].

It should also be mentioned that $\sigma_{\underline{E}}^2$ is not contingent on the number of individuals in a group. Eq. 2 indicates that the expected variance of $\underline{X}$ given random response and $\underline{A}$ = 5 will be 2.0 regardless of group size ($\underline{n_k}$). Moreover, Eq. 2 may be employed to calculate $\sigma_{\underline{E}}^2$ for discrete scales of any length. For example, if $\underline{X}$ assumes values of 1 through 4, then $\sigma_{\underline{E}}^2$ = 1.25, while values of 1 through 7 result in $\sigma_{\underline{E}}^2$ = 4. On the other hand, group size has other implications for the use of Eq. 3, and it is possible to abuse the use of discrete scales. These points are addressed later in this paper.

Within-Group Agreement on Composite Scores

Data employed in an incomplete design are often based on a composite score rather than a single item. Within each group, the composite score takes the form of a sum or a mean per rater over items designed to measure the same construct. Examples would be a set of items to be combined to furnish a composite measure of workgroup morale or team effectiveness in each of $\underline{K}$ groups. We will focus here on an estimate of interrater reliability among raters' composite scores on a set of $\underline{J}$ ($\underline{j}$=1,...,$\underline{J}$) items in each of two groups. It is assumed that (a) the $\underline{J}$ items are a random sample from a well-defined domain of items; (b) the $\underline{n_k}$ raters in each group are randomly sampled from a population of raters, and inferences will be made to that population; and (c) the item variances and interitem covariances are equal, respectively, in the rater population, which implies that the items are considered to be "essentially parallel" indicators of the same construct.

An example of the design in question is presented in Table 2, which represents a facsimile of a problem encountered in research on agreement among performance ratings. The target for Group 1 is a probationary pilot, rated on knowledge of safety procedures independently by five senior pilots ($\underline{n_k}$ = 5) on four items ($\underline{J}$ = 4) designed to measure safety. Each item employs the same seven discrete, equally spaced alternatives ($\underline{A}$ = 7). The target for Group 2 is a different probationary pilot, rated independently by a different set of senior pilots ($\underline{n_k}$ = 6) on the same four safety items. The between-group ICC, based on the rater composite (mean) scores (shown at the bottom of each data matrix) is approximately .00, a result of the fact that the mean composite score for each group is 6.5. Moreover, the within-group *ICC* for each group is approximately .00 [cf. Shrout & Fleiss, 1979, equation for *ICC* (2,1)]. This is a result of the fact that the items, essentially parallel indicators of the same construct, have approximately identical means in each group, from which it follows that each between-item mean square is close to zero.

Insert Table 2 about here

Are we to conclude that the senior pilots lacked agreement in regard to probationary pilots' safety procedures? Certainly not; the variance (now designated $s^2_{X_j}$) and r_{WG} for each item, shown in columns to the right of each data matrix, indicate high levels of agreement. In fact, the average r_{WG}, designated $\overline{r_{WG}}$, is .925 for Group 1 and .93 for Group 2. The separate r_{WG}s were calculated using $\sigma_E{}^2 = 4.0$ [i.e ,7^2-1)/12], and $\overline{r_{WG}}$ may be calculated for each group because the $s^2_{X_j}$ s are the same or similar. An average of the $\overline{r_{WG}}$s for the two groups may also be estimated (≅.93) because of the similar $s^2_{X_j}$ and mean $s^2_{X_j}$, or $\overline{s^2_{X_j}}$, for each group. [$\overline{s^2_{X_j}}$ is equal to the within-item mean square in the ANOVA for each group, and $\overline{r_{WG}}$ may be estimated by $1 - (\overline{s^2_{X_j}}/\sigma_E{}^2)$(Finn, 1970)].

While $\overline{r_{WG}}$ for each group indicates agreement among the raters, it fails to take into account the "boost" in reliability to be expected from combining essentially parallel items to form a composite. That is to say, we would expect the estimate of agreement based on the composite ratings per rater (i.e., a mean or a sum taken over items) to be higher than the estimate of agreement based on the $\overline{r}_{WG}$. This point is illustrated by deriving an equation to estimate agreement among the composite (mean) ratings per rater in each group. [Note that this is *not* the procedure typically employed in *ICC* designs, which consists of estimating interrater reliability among ratee (item) means, based on aggregation over raters.]

The derivation of an interrater reliability coefficient for composite scores in one group is predicated on extrapolating from the logic for one item. The model equation is $\overline{X}_i = \mu + (\overline{X} - \mu) + \overline{e}_i$, where $\overline{X}_i$ and $\overline{e}_i$ are the mean observed and error scores for the ith rater on the J items, respectively, μ is the population mean (equivalent for all items), and $\overline{X}$ is the observed grand mean. As in the case of a single item, variance of the $\overline{X}_i$ scores in a sample arises only from variation in the $\overline{e}_i$. If the J items are equivalent (essentially parallel) and are reflective solely of μ, then the variance of the $\overline{X}_i$ scores will be equal to zero, implying perfect agreement. Variation in the $\overline{X}_i$ scores denotes departure from perfect agreement. Given essentially parallel items, the variance among the $\overline{X}_i$ scores may be estimated by $J(\overline{s^2_{X_j}})/J^2 = \overline{s^2_{X_j}}/J$, where $\overline{s^2_{X_j}}$ is the mean item variance, $J(\overline{s^2_{X_j}})$ estimates the error variance among sums taken over J items (Gulliksen, 1950), and division by J^2 estimates the error variance among means. We will refer to this variance as "error variance".

As discussed earlier, the nonerror or "true variance" for an item is estimated by the heuristic $(\sigma_E^2 - s^2_{X_j})$, where σ_E^2 is employed as a benchmark to indicate the expected variance of the X_i scores on an item j (or X_{ij}) associated with equal likelihood of response and zero interrater reliability. On J essentially parallel items, the true variance for each item may be estimated by $(\sigma_E^2 - \overline{s^2_{X_j}})$, from which it follows that the true variance among means taken over items is estimated by $J^2(\sigma_E^2 - \overline{s^2_{X_j}})/J^2 = (\sigma_E^2 - \overline{s^2_{X_j}})$ [Gulliksen, 1950; where $J^2(\sigma_E^2 - \overline{s^2_{X_j}})$ is the estimated true variance for sums]. Thus, the estimated true variance associated with variation among the $\overline{X}_i$ is the same as that associated with each item.

The interrater reliability associated with agreement among the mean scores in a group, designated $r_{WG(\overline{X}_i)}$, may now be estimated as follows:

$$r_{WG(\overline{X}_i)} = \frac{(\sigma_E^2 - \overline{s^2_{X_j}})}{(\sigma_E^2 - \overline{s^2_{X_j}}) + (\overline{s^2_{X_j}})/J} \qquad (4)$$

$$= \frac{J(\sigma_E^2 - \overline{s^2_{X_j}})}{J(\sigma_E^2 - \overline{s^2_{X_j}}) + \overline{s^2_{X_j}}} \qquad (5)$$

Equations 4 or 5 furnish a computing procedure for $r_{WG(\bar{X}_i)}$. It is also possible to demonstrate that these equations provide an estimate that is equal to the Spearman-Brown (SB) prophecy equation applied to $\overline{r_{WG}}$, the correction factor being the number of items. This equality involves dividing the numerator and denominator of Eq. 5 by σ_E^{2}, which is:

$$r_{WG(\bar{X}_i)} = \frac{J[1 - (\overline{s^2_{X_j}}/\sigma_E^{2})]}{J[1 - (\overline{s^2_{X_j}}/\sigma_E^{2})] + (\overline{s^2_{X_j}}/\sigma_E^{2})} \qquad (6)$$

where $1 - (\overline{s^2_{X_j}}/\sigma_E^{2}) = \overline{r_{WG}}$ and $(\overline{s^2_{X_j}}/\sigma_E^{2}) = 1 - \overline{r_{WG}}$; thus Eq. 6 reduces to $J(\overline{r_{WG}})/[J(\overline{r_{WG}}) + (1 - \overline{r_{WG}})]$, or

$$J(\overline{r_{WG}})/[1 + (J - 1)\overline{r_{WG}}], \qquad (7)$$

which is the SB equation.

Applied to the data in Table 2, Eq. 5 (and Eq. 7) provides the following estimates of agreement for Groups 1 and 2, respectively:

Group 1: $r_{WG(\bar{X}_i)} = 4(4 - .30)/[4(4 - .30) + .30] = .98$

Group 2: $r_{WG(\bar{X}_i)} = 4(4 - .285)/[4(4 - .285) + .285] = .98$

Thus, given that assumptions are satisfied regarding essentially parallel items, $r_{WG(\bar{X}_i)}$ will exceed $\overline{r_{WG}}$, unless the latter statistic is .00 or 1.00. Furthermore, the $r_{WG(\bar{X}_i)}$ may be averaged over the two groups inasmuch as the $\overline{s^2_{X_j}}$ are similar.

Finn (1970) also addressed within-group interrater reliability for a set of items on which item means were essentially equal, and recommended the use of Eq. 7 (the SB equation) to estimate reliability for items, where $\overline{r}_{WG}$ was interpreted as the "mean reliabilityper item." Finn did not, however, furnish statistical justification (or derivation) for the SB equation, including, in particular, the requirement that the items be essentially parallel indicators of the same construct. This is a critical requirement because it justifies the derivation of Eqs. 5 and 7 and suggests that the composite scores are interpretable in reference to an underlying construct. Moreover, the procedures apply to aggregation over items, and not raters, a point confused by Bintig (1980). That is, in a review of the Finn (1970) paper, Bintig interpreted the Finn procedure as an estimator of interrater reliability for aggregates taken over raters for each ratee (items in this paper). It might also be noted that Bintig used an erroneous estimate of σ_E^2 (i.e., a value of 3.5 was used for a seven-point scale, which applies to neither discrete nor continuous scales).

In conclusion $r_{WG(\bar{X}_i)}$ is applicable in incomplete designs when (a) items on which composites (per rater) are based are essentially parallel indicators of the same construct; (b) the mean composite score for each group is approximately the same, and (c) little variation exists among raters in each group. It is possible, of course, for $s_{\bar{X}_j}^2$, and therefore $r_{WG(\bar{X}_i)}$, to vary as a function of group, in which case the $r_{WG(\bar{X}_i)}$ should be interpreted, and reported, separately for each group. On the other hand, the $\overline{s}^2_{X_j}$ and $r_{WG(\bar{X}_i)}$ may be similar over groups, which can be tested by a homogeneity of variance test on the $\overline{s}^2_{X_j}$ (the within-item mean squares). Given similarity, the $r_{WG(\bar{X}_i)}$ can be averaged over groups. [If the decision is to average the $r_{WG(\bar{X}_i)}$ and the n_k differ, there would be little reason to weight the $r_{WG(\bar{X}_i)}$ by n_k because the $r_{WG(\bar{X}_i)}$ are similar. The same argument applies to r_{WG}]. In other words, the reasons for using $r_{WG(\bar{X}_i)}$ rather than an *ICC* approach in incomplete designs are the same as those for using r_{WG}.

Discussion

In regard to incomplete designs, it has been demonstrated that r_{WG} and $r_{WG(\bar{X}_i)}$ provide more accurate estimates of interrater reliability/agreement than an *ICC* when within-group variance is small and differences among group means on an item or a composite (per rater) are essentially nonexistent. In effect, r_{WG} and $r_{WG(\bar{X}_i)}$ furnish an alternative source of estimation when the range of values selected by raters on an item, or on a set of items, is *restricted* for at least some groups. Moreover, unlike the *ICC*, calculation of r_{WG} and $r_{WG(\bar{X}_i)}$ is not dependent on homogeneity of within-group variances, and thus separate estimates of interrater reliability/agreement may be calculated for each group in the absence of such homogeneity. On the other hand, if variances are homogeneous, then the estimates may be averaged over groups to provide an overall estimate of agreement.

Given homogeneity of within-group variances, r_{WG} and $r_{WG(\bar{X}_i)}$ will lose their advantage over the *ICC* for incomplete designs as (a) within-group variances on an item or composite score increase, or (b) the within-group variances remain small but differences among the mean group item/composite scores increase (i.e., *MSB* in Eq. 1 increases in value).[1] The latter point is of major concern because it raises the question of when to use the methods suggested here versus an *ICC* approach, given that at least some variation exists among group means. Future research is needed to answer this question, where, for example, a Monte Carlo study would help to clarify the conditions (e.g., degree of variation among group means, in relation to the magnitudes of $s^2_{X_j}$, $\overline{s^2_{X_j}}$, and n_k) which determine variation among within-group coefficients and *ICC*s in nonobvious situations. A Monte Carlo study is not attempted here, although a brief illustration of point "b" above is presented using the data in Tables 3 and 4. Table 3 has the same pattern of ratings as Table 1 (i.e., low within-group variances); however, a moderate difference in group means (1.05 scale points) was introduced by adding a constant of 1.0 to the scores in Group 2. The resulting *ICC* is .70, which compares much more favorably than the ICC of .00 (Table 1) to the average (over groups) r_{WG} of .88. Table 4 again has the same pattern of ratings as Table 1, but a large difference in group means (2.05 scale points), achieved by adding and subtracting constants. The ICC in Table 4 is .90, which is slightly larger than the average r_{WG} of .88.

Insert Tables 3 and 4 about here

The preceding example is illustrative of the course of action suggested for incomplete designs at the present time. First, if within-group variances appear nonhomogeneous, then conduct a homogeneity of variance test. If homogeneity is rejected, then employ r_{WG} or $r_{WG(\bar{X}_i)}$ to estimate interrater reliability for each group and do not average estimates over groups (at least over nonhomogeneous groups). If homogeneity is not rejected, then compute both an ICC (for an incomplete design) and an average r_{WG} or $r_{WG(\bar{X}_i)}$ over groups. If the estimates differ, then review the raw data matrix and summary statistics (e.g., group means, within-group variances) in relation to the two estimates and ascertain which estimate appears to provide the more accurate point estimate. Finally, report both estimates and the rationale for selecting one as more accurate.

This article is concluded with brief discussion of concerns and potential problems regarding the use of r_{WG} (the discussion applies to $r_{WG(\bar{X}_i)}$). Selvage (1976) and Hsu (1979) argued that the theoretical distribution of X employed in the calculation of σ_E^2 (Eq. 2) should be thought of as normal rather than rectangular. This argument, however, misses the point made earlier that σ_E^2 is a theoretical benchmark used to indicate equal likelihood of response and zero reliability, and makes possible an assessment of the degree to which actual responses resemble random responses. This benchmark is lost if the theoretical distribution is assumed normal for the simple reason that a normal distribution already reflects partial agreement (i.e., there are more scores clustered about the mean than in the tails of the distribtuion). It would appear unwise to employ a theoretical benchmark for lack of agreement that already reflects partial agreement. Consequently, the use of a rectangular distribution for item distributions is recommended.[2]

Selvage (1976, p. 606) argued further that although raters might use only five (or six, etc.) points on a rating (item) scale, the points "are only representative of possible values along the continuum from one to five." This implies that the distribution underlying the random variable X should be regarded as continuous (i.e., represents an infinite number of values) rather than discrete in the calculation of σ_E^2. This argument has validity, but it is also the case that an argument can be made for discrete scales. For example, one could argue that the alternatives in a five to nine point scale encompass sufficiently the degrees (categories) of cognitive differentiation/sensitivity used by most individuals. On the other hand, a continuous scale may be advisable in some cases, and the reader is referred to the Selvage paper for statistical procedures to estimate σ_E^2.

Additional concerns include bias in the estimate of r_{WG}, estimates of less than zero, and artificial manipulation of estimates by unrealistic measurement scales. In regard to bias, r_{WG} may be thought of as a function of two unbiased values; s_X^2 is an unbiased estimate of σ_X^2 for observed values on X, and σ_E^2 is a population parameter. Nevertheless, a ratio of unbiased values (i.e., s_X^2/σ_E^2) is itself biased (Winer, 1971). However, like the ICC in Eq. 1, which is biased for the same reason, the bias in r_{WG} is expected to be minimal for small n_k and essentially negligible for large n_k.

It is possible for r_{WG} to assume values of less than zero. In fact, a number of theoretical distributions of observed ratings results in an s_X^2 greater than σ_E^2, and thus a negative r_{WG}. For example, given one item with $A = 5$ and $n_k = 10$, if five raters selected alternative 1 and five raters selected alternative 5, r_{WG} would be equal to -1.22 (i.e., $1 - \frac{4.44}{2.0}$). However,every distribution of observed X that could result in a negative r_{WG} would reflect rather serious degrees of disagreement. Consequently, it is recommended that negative estimates be set equal to zero to indicate lack of agreement among raters.

Of special importance is the fact that it is possible to manipulate artificially the value of r_{WG} by constructing unrealistic measurement scales. Suppose, for example, that an individual constructed a meaningful seven-point scale that encompassed all likely responses. Suppose further that this individual added three spurious alternatives to each end of the scale; that is, alternatives with a zero base-rate (e.g., a teacher evaluation such as: This teacher has never made even the most trivial mistake). We now have a 13-point scale, resulting in an $\underline{\sigma_E}^2$ = 14, when in fact the true scale with seven points should have a $\underline{\sigma_E}^2$ = 4. Finally, suppose that the distribution of observed values on $\underline{X}$ is uniform on the true seven-point scale, which suggests an interrater reliability of zero. The interrater reliability is not, however, zero. For example, with $\underline{n_k}$ = 21 and $\underline{s_X}^2$ = 4.2, rather than the accurate $\underline{r_{WG}} \approxeq .00$ [i.e., 1 - (4.20/4.0], $\underline{r_{WG}}$ is .70 [i.e., 1 - 4.20/14.0). This is the result of artificially adding six spurious alternatives to the scale.

A different problem occurs with using too short a scale, where, for example, the observed distribution on a three-point scale could appear approximately uniform. On a longer but meaningful scale (e.g., seven points), the scores might spread, but the locus of points could remain dense within the original three points. These conditions imply that the $\underline{r_{WG}}$ for the three-point scale will be artificially low. In general, it would seem that too short a scale would be a result of poor research practice rather than artificial manipulation, although the latter condition is a possibility if a vested interest existed in obtaining a low interrater reliability.

In conclusion, use of the procedures discussed here rests on the assumption that the measurement scale is meaningful. This does not suggest that all points on the scale have to be used in every sample; it suggests only that the scale is sensitive to, and limited to, psychometrically reliable differentiation on the measured attribute. Valid scaling procedures in conjunction with professional and ethical judgment should satisfy this criterion.

References

Bartko, J. J. On various intraclass correlation reliability coefficients. Psychological Bulletin, 1976, 83, 762-765.

Bintig, A. The efficiency of various estimations of reliability of rating scales. Educational and Psychological Measurement, 1980, 40, 619-643.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cooper, M. An exact probability test for use with Likert-type scales. Educational and Psychological Measurement, 1976, 36, 647-655.

Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Finn, R. H. A note on estimating the reliability of categorical data. Educational and Psychological Measurement, 1970, 30, 71-76.

Fleiss, J. L., & Shrout, P. E. Approximate interval estimation for a certain intraclass correlation coefficient. Psychometrika, 1978, 43, 259-262.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Hsu, L. Agreement or disagreement of a set of Likert-type ratings. Educational and Psychological Measurement, 1979, 39, 291-295.

Insel, P. M., & Moos, R. H. Psychological environments: Expanding the scope of human ecology. American Psychologist, 1974, 29, 179-188.

James, L. R., & Jones, A. P. Organizational climate: A review of theory and research. Psychological Bulletin, 1974, 81, 1096-1112.

Kraemer, H. C., & Korner, A. F. Statistical alternatives in assessing reliability, consistency, and individual differences for quantitative measures: Application to behavioral measures of neonates. Psychological Bulletin, 1976, 83, 914-921.

Mood, A. M., Graybill, F. A., & Boes, D. C. Introduction to the theory of statistics. New York: McGraw-Hill, 1974.

Saal, F. E., Downey, R. G., & Lahey, M. A. Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 1980, 88, 413-428.

Selvage, R. Comments on the analysis of variance strategy for computation of intraclass reliability. Educational and Psychological Measurement, 1976, 36, 605-609.

Shrout, P. E., & Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 1979, 86, 420-428.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1971.

Footnotes

Support for this project was provided under Office of Naval Research Contract N00014-80-C-0315, Office of Naval Research Project NR170-904. Opinions expressed are those of the authors and are not to be construed as necessarily reflecting the official view or endorsement of the Department of the Navy.

The authors wish to thank John R. Bruni, Dennis M. Irons, George W. Joe, Charles V. Riche, S. B. Sells, and C. Michael York for their helpful suggestions and advice.

Requests for reprints should be sent to Lawrence R. James, School of Psychology, Georgia Institute of Technology, Atlanta, Georgia 30332.

[1] All references to the *ICC* approach in this section refer to the *ICC* for incomplete designs (Eq. 1). It is assumed that lack of variation among item means would generally preclude the use of the within-group *ICC*.

[2] The underlying theoretical distribution for rater composite scores in Eqs. 5 and 7 is normal, a result of the central limit theorem. This does not detract from the fact that equal likelihood of response on each item implies that one begin with a rectangular distribution for each item.

Table 1

Intraclass Correlation for Twenty Raters

Nested in Each of Two Groups

Scale for Variable *X*	Frequencies of Scores in Group 1	Frequencies of Scores in Group 2
1	0	0
2	2	2
3	16	15
4	2	3
5	0	0
Mean:	3.00	3.05
Variance:	.211	.261

Analysis of Variance

Source	*df*	*SS*	*MS*	
Between-Groups	1	.025	.025	*F* = .106*NS*
Within-Groups	38	8.959	.236	

Intraclass Correlation

$$\underline{ICC} = \frac{.025 - .236}{.025 + (19)(.236)}$$

$$= -.047$$

$$\cong .00$$

Note: *NS* = not significant at $p < .05$.

Table 2

Between-Group and Within-Group Intraclass Correlations

For Two Groups

Group 1

Item	Rater 1	2	3	4	5	Mean	$s^2_{X_j}$	r_{WG}
1	7	6	7	6	7	6.6	.30	.92
2	6	6	7	7	6	6.4	.30	.92
3	6	6	7	6	7	6.4	.30	.92
4	7	7	7	6	6	6.6	.30	.92
Mean	6.5	6.25	7.0	6.25	6.5			

Group 2

Item	Rater 1	2	3	4	5	6	Mean	$s^2_{X_j}$	r_{WG}
1	6	6	7	7	7	7	6.67	.27	.93
2	7	6	6	7	6	6	6.30	.27	.93
3	7	7	7	6	6	6	6.5	.30	.92
4	6	7	6	7	6	7	6.5	.30	.92
Mean	6.5	6.5	6.5	6.75	6.25	6.50			

Within-Group ICC

Source	df	SS	MS
Between-Item	3	.195	.065
Within-Item	16	4.80	.30
Between Rater	4	1.50	.375
Residual	12	3.30	.275

ICC ≅ .00[1]

Between-Group ICC

Source	df	SS[2]	MS
Between-Groups	1	0	0
Within-Groups	9	.501	.056

ICC ≅ .00

Within-Group ICC

Source	df	SS	MS
Between-Item	3	.411	.137
Within-Item	20	5.70	.285
Between Rater	5	.50	.10
Residual	15	5.20	.347

ICC ≅ .00[1]

1/ Based on Shrout and Fleiss (1979) Equation for ICC (2,1).

2/ Computations based on statistical equations.

Table 3

Comparison of Interrater Reliabilities
for Two Groups with Moderate Mean Differences

Scale for Variable X	Frequencies of Scores in Group 1		Frequencies of Scores in Group 2
1	0		0
2	2		0
3	16		2
4	2		15
5	0		3
Mean	3.00		4.05
Variance	.211		.261
r_{WG}[1]	.89		.87
ICC		.70	

Analysis of Variance

Source	df	SS	MS	
Between-Groups	1	11.025	11.025	F=46.79*
Within-Groups	38	8.959	.236	

* $p < .01$

[1] Interrater reliability based on $1-(s_X^2/\sigma_E^2)$, where $\sigma_E^2 = 2.0$.

Table 4

Comparison of Interrater Reliabilities for Two Groups with Large Mean Differences

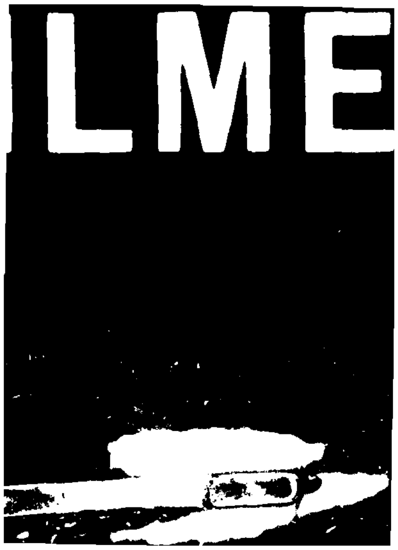
Scale for Variable $\underline{X}$	Frequencies of Scores in Group 1		Frequencies of Scores in Group 2
1	2		0
2	16		0
3	2		2
4	0		15
5	0		3
Mean	2.00		4.05
Variance	.211		.261
$\underline{r_{WG}}$[1]	.89		.87
$\underline{ICC}$		.90	

Analysis of Variance

Source	df	SS	MS	
Between-Groups	1	42.025	42.025	$\underline{F}$=178.43*
Within-Groups	38	8.959	.236	

* $\underline{p} < .01$

[1] Interrater reliability based on 1 - ($\underline{s_X}^2/\sigma_{\underline{E}}^2$), where $\sigma_{\underline{E}}^2$ = 2.0.


LME